Research Statement
Hyrum D. Carroll, PhD

Research Interests

Computational biology is an area of research at the intersection of computer science and biology. This is my research area because I get to leverage high performance computing to answer computational problems for molecular biology. As a doctoral student and a postdoctoral researcher, I have worked on a variety of projects including both pairwise and multiple sequence alignment, phylogeny search, and biological networks. My future work will continue along these lines with an expansion into leveraging high-throughput sequencing (also called next-generation sequencing).

Past Research

- **Constraint Based Genetic Multiple Sequence Alignment** (Carroll *et al.*, 2008, 2009)
  Traditional approaches to multiple sequence alignment (MSA) of genetic data sets use empirically derived substitution matrices focused on the mutation frequencies of amino acids. Typically ignored is the fact that the physicochemical properties (e.g., hydrophobicity, polarity, volume, etc.) of residues play an important role in the fitness of substituting one residue for another. As part of my dissertation, this work focused on answering a few questions related to MSAs to develop a new MSA algorithm. These questions include:

  > What are the physicochemical properties that are most important to MSAs?
  >
  > What physicochemical properties do existing substitution matrices (e.g., BLOSUM, GONNET, and PAM) favor?
  >
  > If a MSA algorithm used different substitution matrices for different secondary structure assignments (i.e., α-helices, β-sheets and "loop"), how much does that improve its accuracy (as compared to a static substitution matrix)?
  >
  > What is the difference in accuracy of a MSA algorithm if secondary structure and conserved regions are used as constraints?

  These questions ultimately led to incorporating secondary structure elements into a new multiple sequence alignment algorithm called ChemAlign. ChemAlign uses the secondary structure both as a constraint and to determine which substitution matrix to use. In addition to the value from the respective substitution matrix during the typical Needleman-Wunsch portion of the alignment algorithm, the weights of all applicable constraints are added to determine if the characters should be aligned or a gap should be introduced. To illustrate this concept, assume we're aligning the two sequences in Figure 1. Aligning two characters that are both part of homologous β-sheets receive an additional weight proportional to the weight associated with each β-sheet. Similarly, conserved regions (as determined by local alignments) are calculated and incorporated with weights proportional to their respective quality. ChemAlign is able to produce more biologically correct alignments than commonly used algorithms.

- **DNA Reference Alignment Database** (Carroll *et al.*, 2007, 2008)
  Multiple sequence alignments (MSAs) are at the heart of bioinformatics analysis. Recently, a number of multiple protein sequence alignment benchmarks (i.e., BAliBASE, OXBench, PREFAB and SMART) have been released to evaluate new and existing MSA applications. These databases have been well received by researchers and help to quantitatively evaluate MSA programs on protein sequences. Unfortunately, an analogous DNA benchmark was not available, making evaluation of MSA programs

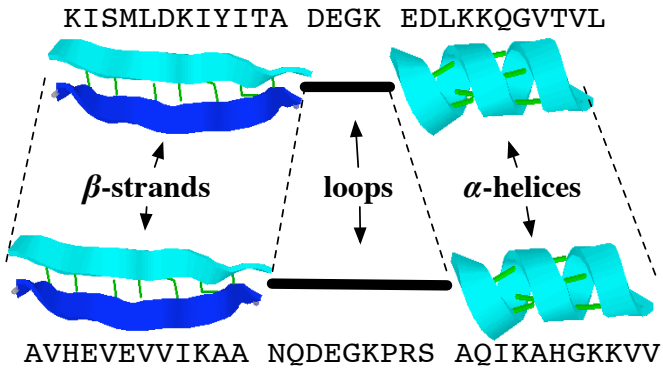

Figure 1: Graphical representation of a constraint alignment of two sequences. Here, each sequence has a β-sheet, a "loop" (indicated with a solid thick line) and then an α-helix. Regions marked by the dotted lines should be aligned together by adding gaps.

difficult for DNA sequences. This work is the first multiple DNA sequence alignment benchmark that is 1) comprised of protein-coding portions of DNA 2) based on biological features such as the tertiary structure of encoded proteins. These reference DNA databases contain a total of 3,545 alignments, comprising of 68,581 sequences. Two versions of the database are available: `mdsa_100s` and `mdsa_all`. The `mdsa_100s` version contains the alignments of the first TBLASTN hit for each protein sequence that had a 100% sequence identity match. The `mdsa_all` version includes the first hit with an E-value score above the threshold of 0.001 for each protein sequence. These databases have been well received by researchers evaluating MSA programs on DNA sequences. The databases and a case study using the databases are publicly available at **http://csl.cs.byu.edu/mdsas/**.

- **Biological Network Paths in Down Syndrome Mice** (Rungta *et al.*, 2007; Clement *et al.*, 2009) Biologists attribute Down Syndrome characteristics to the interactions of chromosome 21 genes with non-chromosome 21 genes. The size and complexity of gene regulatory networks make it hard for biologists to empirically determine how chromosome 21 genes are related to non-chromosome 21 genes. We propose using novel paths by linking genes that exist in multiple pathways. Furthermore, additional gene relationships can be added by mining PubMed abstracts. Combining these sources of data allows for analysis of paths between genes to suggest novel relationships. We are explicitly looking at genes associated with chromosome 21 genes to suggest a list of interesting gene interactions that have not been previously considered. A web server version of our query tool is publicly available at **http://dna.cs.byu.edu/pathgen/**. Researchers can use this tool to generate inter-pathway paths between two sets of genes. Results are graphically displayed with hyperlinked references.

Current Research

- **TAP: An Evaluation Measure for Retrieval** (Carroll *et al.*, 2010a, 2010b)
Retrieval is a common task (e.g., database, biomedical text, and web searches). In many fields, e.g., bioinformatics, a threshold is often applied to the retrieved results. Current homology search studies often use pooled receiver operating characteristic (ROC) curves to evaluate the performance of different methods. Unfortunately, this method does not accurately reflect the actual usage of the retrieval programs. Furthermore, it is susceptible to being arbitrarily skewed by a single query. To remedy these issues, we propose the Threshold Average Precision (TAP), a method based on average precision (from information retrieval) that explicitly incorporates E-values. Along with this method, we propose additional criteria for ideal retrieval methods. The TAP measure is a robust method that addresses these criteria. Our TAP web server is available at `http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html.ncbi/tap/tap.cgi`. This criterion has been well received by researchers, for example, it was recently adopted as the single evaluation measure for a task at the BioCreAtIvE III workshop.

- **PSI-GLOBAL: Iterative Database Searching Using Multiple Contiguous Strings**
  Every second of every day, researchers use homology search computer programs to find sequences that are evolutionarily related to a query sequence. They utilize these results for genome annotation, 2D and 3D structure prediction, to determine evolutionary relationships between organisms among other uses. Exemplary studies that use these results include discovering the origins of flu outbreaks and forensic investigations.

  Most homology search algorithms find "local" alignments. On the other hand, using "semi-global" alignment can provide more accurate statistical estimators. To take advantage of the inherent domain structures in sequences with a semi-global approach, we developed PSI-GLOBAL, an algorithm that uses multiple contiguous strings in the query sequences to look for similar sequences in a database. Preliminary results indicate that PSI-GLOBAL's statistical score is more accurate than the commonly used "local" similarity search algorithms (e.g., RPS-BLAST (Schäffer *et al.*, 1999) and HMMER (http://hmmer.janelia.org)). To make execution time for PSI-GLOBAL practical, we leverage the search and database heuristics from the BLAST framework (Camacho *et al.*, 2009).

Future Research

- **False Discovery Rates for Genetic Sequence Database Retrieval**
  Genetic sequence database search algorithms usually calculate the probability of a match between a query sequence and a sequence in the database as an E-value, an estimate of the number of matches with an equal or better score that would occur by chance in a particular database. E-values are only comparable for the same database. This can be problematic in bioinformatics because sequence databases are often updated daily or weekly. Furthermore, scores are harder to interpret when specialty or subset databases are created. To address these issues, we propose using an estimate of the false discovery rate (FDR) as the statistic for search algorithms. The FDR can be estimated by the P-value that is ubiquitous in search algorithms. It is comparable across databases and is therefore more intuitive for users, leading to more accurate usage.

- **Multiple Protein Domain Prediction**
  A domain is a contiguous string of amino acid characters. Substructures of both a sequence and its 3D configuration are comprised of one or more domains. Knowing the locations of domains is very useful for calculating multiple sequence alignments, prediction of the 3D structure among other uses in bioinformatics. Given PSI-GLOBAL's inherent representation of blocks, it can be utilized to predict domain locations. PSI-GLOBAL is typically used to learn about a query sequence by finding sequences in a database that are biologically similar. Often the entries in the database have more information that can be applied to the query. To predict domain locations a reverse strategy can be applied; a set of sequences with known domain information searching against a database comprised solely of the query sequence. The locations of high scoring matches indicate probable domain boundaries.

Summary

In my opinion, interdisciplinary work is where most of the new contributions to science will be. The areas of computational biology and bioinformatics blend computing into biology. I have worked on a wide range of projects in this area from multiple sequence alignment to biological networks. Additionally, my postdoctoral research has focused on algorithmic development and evaluation methods for pairwise sequence alignment. My roadmap for future research continues to be based on pairwise sequence alignment with the additional of incorporating solutions that involve high-throughput sequencing.

References

Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC bioinformatics*, **10**:421.

Carroll,H., Beckstead,W., O'Connor,T., Ebbert,M., Clement,M., Snell,Q. and McClellan,D. (2007) DNA Reference Alignment Benchmarks Based on Tertiary Structure of Encoded Proteins. *Bioinformatics*, **23**, pp 2648-2649. (web, pdf)

Carroll,H.D. (2008) Biologically Relevant Multiple Sequence Alignment. Ph.D. dissertation, Brigham Young University. (pdf)

Carroll,H., Clement,M., Snell,Q. and McClellan,D. (2009) ChemAlign: Biologically Relevant Multiple Sequence Alignment Using Physicochemical Properties. *Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine*, pp 70–73. (pdf)

Carroll,H.D., Kann,M.G., Sheetlin,S.L. and Spouge,J.L. (2010a) Threshold Average Precision (TAP-k): A Measure of Retrieval Efficacy Designed for Bioinformatics. *Bioinformatics*, **26**, pp 1708–1713. (web, pdf)

Carroll,H.D., Kann,M.G., Sheetlin,S.L. and Spouge,J.L. (2010b) Threshold Average Precision (TAP-k): A Retrieval Efficacy Measure for Bioinformatics. *Intelligent Systems for Molecular Biology*. (web, pdf)

Clement,K., Gustafson,N., Berbert,A., **Carroll,H.**, Merris,C., Olsen,A., Clement,M., Snell,Q., Allen,J. and Roper,R.J. (2010) PathGen: A Transitive Gene Pathway Generator. *Bioinformatics*, **26**, pp 423–425. (web, pdf)

Rungta,N., **Carroll,H.**, Mercer,E., Roper,R., Clement,M. and Snell,Q. (2007) Analyzing Gene Relationships for Down Syndrome with Labeled Transition Graphs. *Proceedings of Formal Methods in Computer Aided Design*, pp 216-219. (web, pdf)

Schäffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, pp 1000–1011.